

An Analysis of Simple Data Augmentation for Named Entity Recognition



Xiang Dai^{♣♣♦} and Heike Adel[♣]
 ♣ Bosch Center for Artificial Intelligence ♠ University of Sydney ♦ CSIRO Data61



Motivation

- Data augmentation: expanding the training set by applying transformations to original training instances without changing their labels.
- Existing augmentations on sentence-level NLP tasks create augmented instances by **manipulating a few words** in the original instance, such as word replacement, random deletion or word position swap.
- We design **simple** data augmentation methods for NER, which **do not rely on any externally trained models**.

Takeaways

- Simple data augmentations can improve over the baseline where no augmentation is used. Synonym replacement outperforms other augmentations on average when BERTs are used, whereas mention replacement appears to be most effective for recurrent models.
- Applying multiple data augmentation methods together outperforms any single data augmentation on average.
- Data augmentations are more effective when the training sets are small.

Code on Github

- <https://bit.ly/32xeK0u>



Data Augmentation for NER

- Label-wise token replacement (LwTR): randomly decide whether each token should be replaced. If yes, randomly select another token from the original training set with the same tag as the replacement.
- Synonym replacement (SR): similar to LwTR, except that the replacement is one of the synonyms retrieved from WordNet.
- Mention replacement (MR): randomly decide whether each mention should be replaced. If yes, randomly select another mention from the original training set which has the same entity type as the replacement.
- Shuffle within segments (SiS): split the token sequence into segments of the same label. Then for each segment, randomly decide whether it should be shuffled. If yes, the order of the tokens within the segment is shuffled.
- All: for each training instance, create multiple augmented instances, one per aforementioned augmentation method.

	Instance												
None	She	did	not	complain	of	headache	or	any	other	neurological	symptoms	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	
LwTR	L.	One	not	complain	of	headache	he	any	interatrial	neurological	current	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	
SR	She	did	non	complain	of	headache	or	whatsoever	former	neurologic	symptom	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	
MR	She	did	not	complain	of	neuropathic	pain	syndrome	or	acute	pulmonary	disease	.
	O	O	O	O	O	B-problem	I-problem	I-problem	O	B-problem	I-problem	I-problem	O
SiS	not	complain	She	did	of	headache	or	neurological	any	symptoms	other	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	

Evaluation Results

- MaSciP: synthesis procedures annotated with synthesis operations and their typed arguments (e.g., Material, Synthesis-Apparatus, etc.).
- i2b2-2010: identifying Problem, Treatment and Test from patient reports.

Model	Method	MaSciP				i2b2-2010				Δ
		S	M	L	F	S	M	L	F	
Recurrent	No augmentation	53.0 _{±3.2}	63.0 _{±0.6}	70.3 _{±0.8}	76.4 _{±0.4}	17.1 _{±2.0}	43.3 _{±1.2}	54.1 _{±0.6}	81.1 _{±0.2}	
	Label-wise token rep.	59.7 _{±0.6}	65.5 _{±0.6}	71.4 _{±0.4}	76.3 _{±0.8}	26.7 _{±0.8}	44.3 _{±0.8}	54.5 _{±0.8}	81.0 _{±0.2}	2.6
	Synonym replacement	60.1 _{±0.5}	65.4 _{±0.4}	70.8 _{±0.6}	76.7 _{±0.8}	25.9 _{±0.5}	44.1 _{±0.5}	54.4 _{±1.5}	81.0 _{±0.3}	2.5
	Mention replacement	60.6 _{±0.6}	65.4 _{±0.4}	71.9 _{±0.5}	76.0 _{±0.8}	25.9 _{±0.7}	45.5 _{±0.4}	55.0 _{±0.2}	81.4 _{±0.2}	2.9
	Shuffle within segments	58.8 _{±0.7}	64.6 _{±0.4}	70.5 _{±0.8}	77.1 _{±0.3}	25.2 _{±0.5}	44.4 _{±0.6}	53.5 _{±0.9}	80.6 _{±0.3}	2.0
	All	60.8 _{±1.3}	67.0 _{±0.8}	72.1 _{±0.7}	76.6 _{±0.4}	26.9 _{±0.7}	45.4 _{±0.6}	54.6 _{±0.9}	81.5 _{±0.2}	3.3
Transformer	No augmentation	68.1 _{±0.6}	72.7 _{±0.3}	77.3 _{±0.5}	79.8 _{±0.7}	35.1 _{±1.1}	62.7 _{±1.5}	70.2 _{±0.3}	87.8 _{±0.2}	
	Label-wise token rep.	70.0 _{±0.8}	72.8 _{±0.2}	76.0 _{±0.6}	80.2 _{±0.6}	39.3 _{±1.7}	64.8 _{±1.3}	71.2 _{±0.4}	87.5 _{±0.2}	1.0
	Synonym replacement	70.6 _{±1.2}	73.9 _{±0.1}	76.8 _{±0.4}	79.7 _{±0.5}	42.3 _{±1.3}	65.3 _{±0.3}	70.5 _{±2.3}	87.7 _{±0.4}	1.6
	Mention replacement	70.5 _{±0.8}	73.3 _{±0.4}	76.7 _{±0.7}	80.0 _{±0.3}	40.1 _{±2.5}	64.2 _{±1.2}	70.8 _{±0.7}	87.8 _{±0.2}	1.2
	Shuffle within segments	70.5 _{±0.4}	73.1 _{±0.6}	76.7 _{±0.3}	80.3 _{±0.5}	39.4 _{±1.6}	63.9 _{±1.4}	71.2 _{±1.2}	87.7 _{±0.2}	1.1
	All	71.2 _{±0.8}	73.1 _{±0.6}	76.9 _{±0.4}	80.5 _{±0.4}	41.5 _{±0.9}	65.2 _{±0.3}	72.3 _{±1.3}	87.2 _{±0.3}	1.8

Δ column shows the averaged improvement due to data augmentation. **S, M, L** contain 50, 150, 500 training instances, respectively; **F** uses the complete training set.